

Analisi PCA

ELABORAZIONE DEI DATI SCIENTIFICI

[262-009]

Autore:

Testoni Manuel (176931)

November 2, 2025

Contents

1	Introduzione	5
1.1	Obiettivo	5
1.2	Descrizione del Dataset	5
2	Risultati	7
2.1	Scelta del Pre Processing	7
2.2	Scelta delle PC's	7
2.3	Grafici	7
2.3.1	Loadings	8
2.3.2	Scores Plot	9
2.3.3	PC's	10
2.3.4	Residui	11
3	Conclusioni	13

1

Introduzione

1.1 Obiettivo

In questo studio si prende in analisi il dataset "Wines" utilizzando tecniche di riduzione di dimensionalità come PCA. Si è preso in considerazione l'utilizzo di PCA in quanto ci permette di descrivere in modo sintetico la struttura multivariata del dataset. Questo è specialmente utile quando le molteplici variabili sono correlate tra loro. Riducendo la dimensionalità del dataset da 14 variabili a solamente 2 o 3, chiamandole componenti principali, siamo in grado di mantenere un alta percentuale di informazione (varianza), perdendone poca, rendendo più facili da confrontare le tre categorie di vini. Utilizzare PCA ci permette di estrarre e visualizzare la struttura del dataset, rendendo possibile identificare pattern, distribuzioni e raggruppamenti fra le denominazioni.

1.2 Descrizione del Dataset

Il dataset analizzato è costituito da campioni di vino rosso appartenenti a tre denominazioni differenti: Barolo (OLO), Barbera (ERA) e Grignolino (GR). Per ogni campione sono stati misurati 14 descrittori chimico-fisici che caratterizzano la composizione del vino. La maggior parte delle variabili riguarda la presenza di composti fenolici. Sono inoltre presenti variabili relative ad acidi organici, componenti minerali, caratteristiche cromatiche e indici spettrali UV-Vis. Data la correlazione tra molte di queste variabili (per esempio quelle relative ai composti fenolici), essendo in un contesto multivariato a dimensionalità "elevata" per effettuare un'analisi "ad occhio", è necessario l'impiego di tecniche di riduzione della dimensionalità.

Vediamo di seguito la tabella riportante la descrizione per ogni variabile del nostro dataset.

Variabile	Descrizione
Alcohol	grado alcolico (% in volume)
Malic.ac	contenuto acido malico, uno dei principali acidi organici presenti nel vino
Ash	ceneri (parte inorganica / sali)
Alcalinity.of.ash	alcalinità delle ceneri, parametro di controllo del pH
Mg	contenuto di Magnesio (g/kg)
Phenols	contenuto totale di composti fenolici: sostanze naturali che danno colore e contributi organolettici
Flavanoids	composti flavonoidi (principali polifenoli)
Nonflav.phen	composti fenolici non-flavonoidi che conferiscono caratteristiche specifiche al vino
Proanthoc	contenuto totale di proto-antocianine (fenoli antiossidanti del vino rosso)
Color.intensity	intensità del colore rosso
Hue	saturazione della tinta (colore)
OD280/OD315	rapporto fra 2 lunghezze d'onda nell'UV-Vis (indicatore qualità fenolica)
Proline	prolina (mg/L), amminoacido caratteristico del vino

Table 1.1: descrizione delle variabili del dataset dei vini

2

Risultati

2.1 Scelta del Pre Processing

In questo dataset, dove le variabili hanno grandezze fisiche eterogenee, non possiamo affidarci alla varianza assoluta. È necessario rimuovere l'effetto della scala per rendere le variabili comparabili e questo possiamo farlo con Autoscaling. Questo è necessario in quanto PCA, è una tecnica che si basa sulla varianza e andando ad utilizzare mean centering andremmo semplicemente a spostare le variabili attorno la media 0 senza uniformarne la varianza e ottenendo che le variabili avrebbero un peso diverso. In questo caso se avessimo usato mean centering avremmo trovato che la nostra variabile "prolina" ha per esempio un range di valori tra circa 0.300 e circa 1300, mentre invece l'OD Ratio ha un range da circa 0.9 e 1.3 circa, questo vorrebbe dire che PCA per spiegare la varianza generale prenderebbe sempre la variabile prolina come componente, ma questo è solamente un artefatto dato dal valore numerico alto della variabile. Andando ad utilizzare autoscaling ci assicuriamo di dare la stessa importanza a tutte le variabili.

2.2 Scelta delle PC's

La selezione del numero di componenti principali non è stata fatta "visivamente", ma seguendo criteri standard in PCA. Sono stati considerati: la percentuale di varianza cumulativa spiegata e il "gomito" (elbow) nella scree plot degli eigenvalues.

La PCA dopo autoscaling mostra che le prime 3 componenti spiegano complessivamente circa 66% della varianza totale e dopo questa soglia la scree plot evidenzia un chiaro "punto di gomito", quindi le PC successive non catturano una varianza tale da renderle informative. Di conseguenza sono state considerate solo le prime 3 componenti principali.

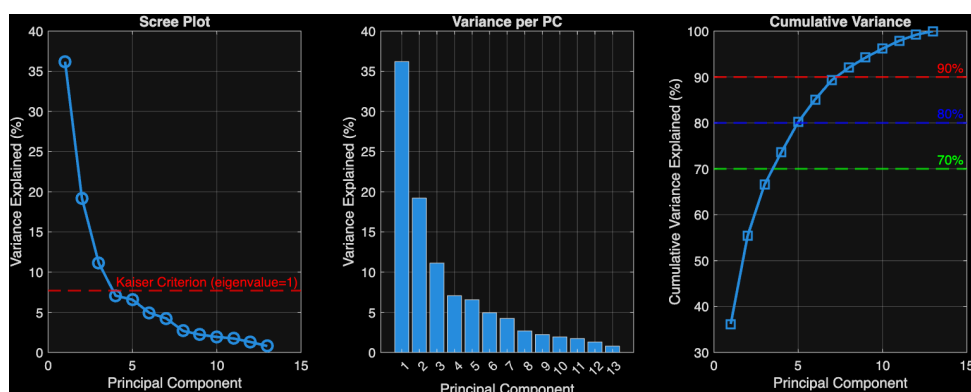


Figure 2.1: Grafici elbow, varianza cumulativa spiegata dalle PC's

2.3 Grafici

2.3.1 Loadings

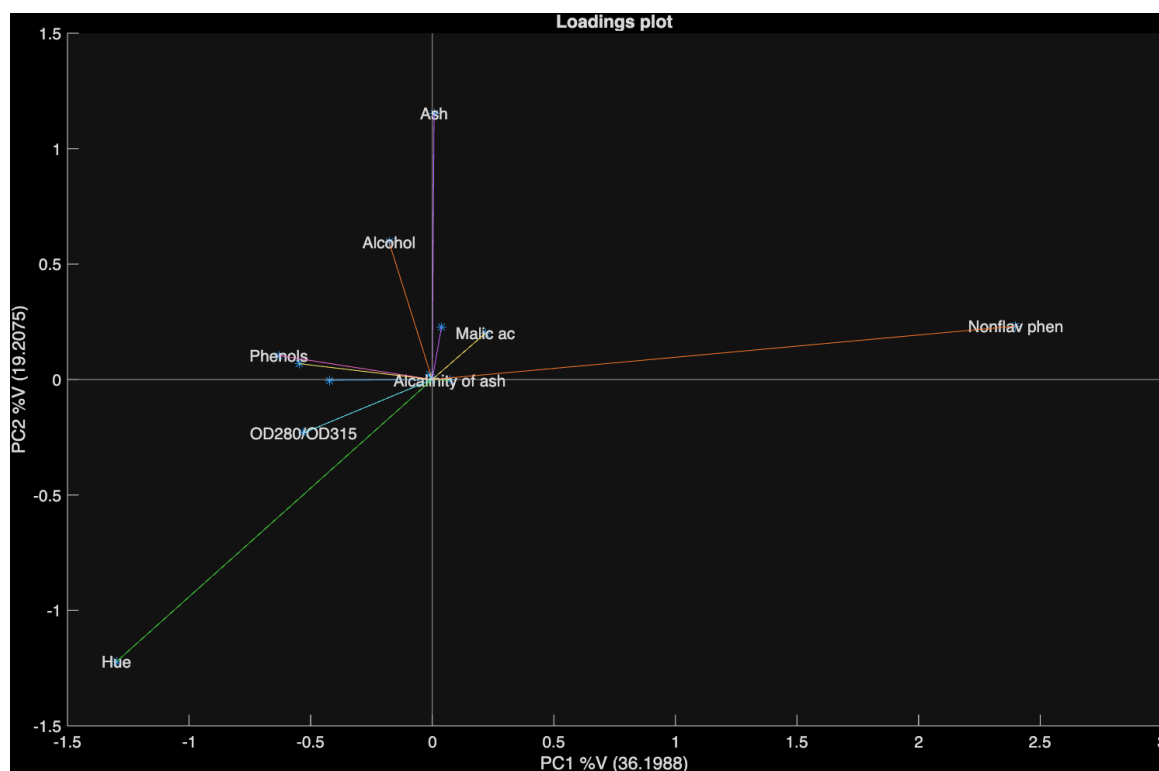


Figure 2.2: Grafico dei Loadings

Il loadings plot rappresenta le variabili nello spazio delle componenti principali e come queste vi contribuiscono. Il contributo lo determiniamo in base al modulo e al verso del vettore tracciato per ogni variabile nel grafico qua sopra. Osservandolo possiamo determinare che la variabile che contribuisce di più in verso positivo per PC1 è la Nonflav phen mentre Hue contribuisce ad entrambe in modo fortemente negativo. Siamo in grado di dire che essendo queste due variabili lontane e in verso opposto nello spazio delle componenti principali queste due variabili sono anti correlate. Altre variabili come le ceneri o l'alcool danno un contributo medio, mentre i fenoli e acidi vari contribuiscono in quantità minori.

2.3.2 Scores Plot

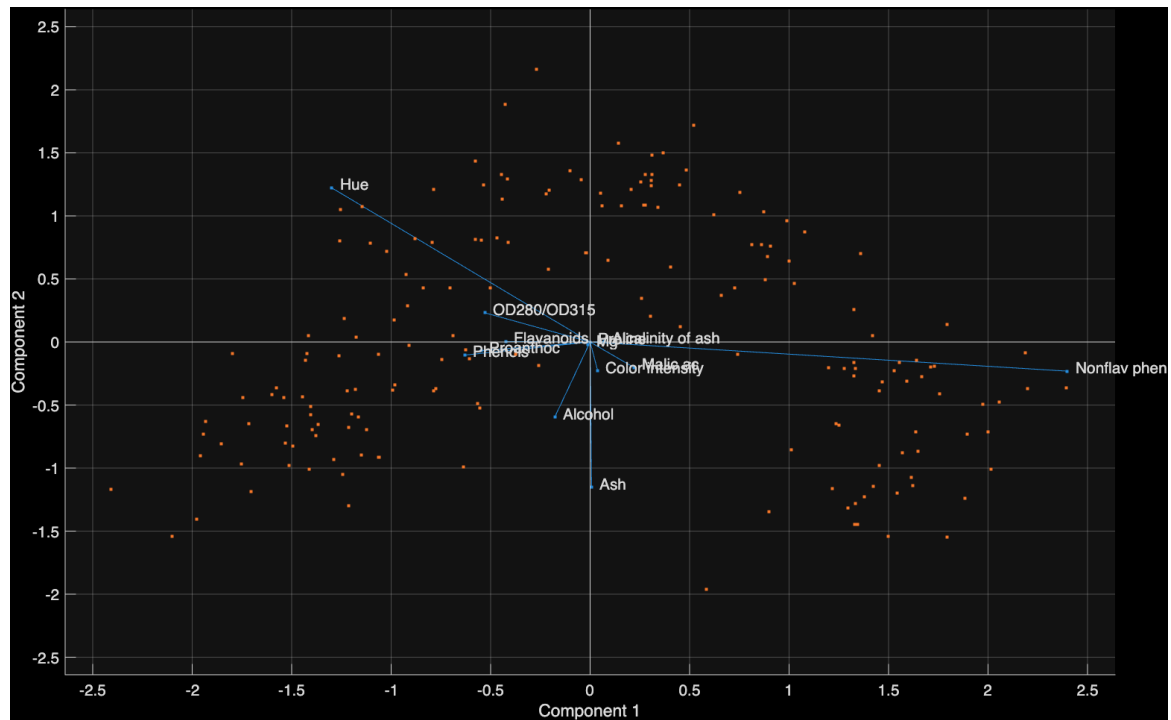


Figure 2.3: Grafico degli scores

Lo score plot mostra la proiezione dei campioni nello spazio PC1–PC2. I campioni con coordinate simili formano gruppi. In generale guardando questo grafico possiamo dire che un gruppo di vini si orienta nello spazio delle componenti principali verso la nonflav phen ovvero sono più ricchi di non flavonoidi, mentre un altro gruppo si allontana dalla variabile "Hue", che rappresenta la composizione cromatica dei vini. Un altro gruppo è individuabile nella zona relativa a valori positivi di PC2 e valori nell'intervallo di -0.5,0.5 per quanto riguarda PC1.

2.3.3 PC's

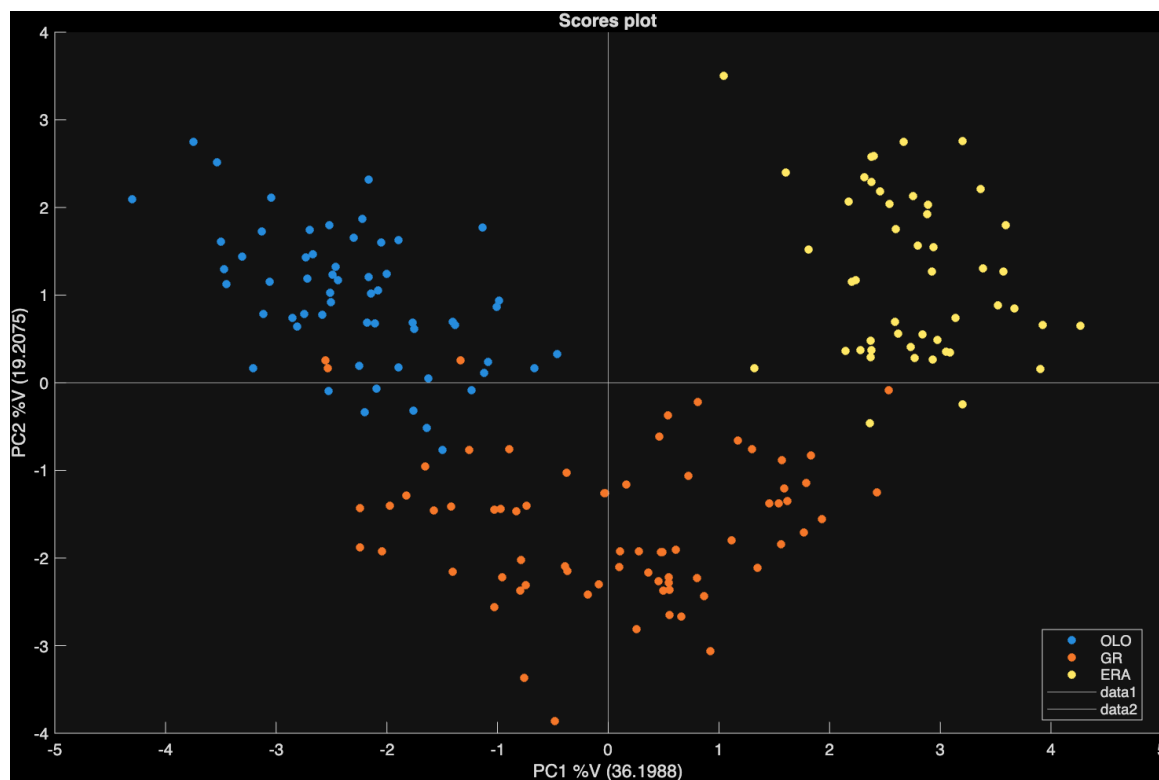


Figure 2.4: Scatter plot con PC's

Con questo grafico siamo invece in grado di determinare la distribuzione delle 3 varie denominazioni di vini all'interno del nostro spazio delle componenti principali. Vediamo dunque che i tre gruppi identificati prima si traducono proprio in 3 cluster rispettivamente di ogni famiglia di vini. Vi sono dunque tre campioni di Grignolino che, come possiamo vedere in figura, si discostano dalla distribuzione degli altri campioni e si avvicinano a quella dei vini "Barolo" e un campione che si avvicina invece alla famiglia dei vini "Barbera". Possiamo dunque dire che generalmente, vini della stessa tipologia ma prodotti in cantine diverse conservano più o meno i valori per le variabili prese in considerazione in questo dataset, andando così a definire delle regioni nello spazio delle componenti principali dove i campioni di ogni famiglia di vini si raggruppa in clusters.

Siamo in grado di dire inoltre come i vini "Barolo" siano distinti fortemente dalla variabile Nonflav Phen e come invece i vini di tipo "Barbera" presentino valori positivi della variabile Hue.

2.3.4 Residui

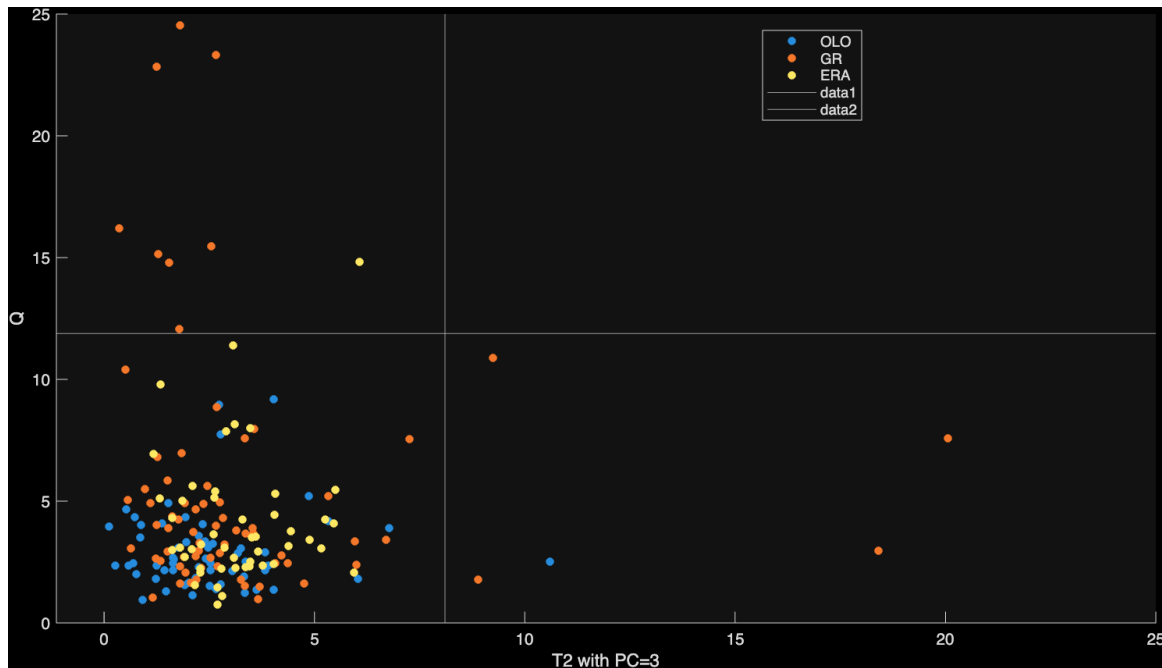


Figure 2.5: Grafico dei residui

Analizzando il grafico T^2 vs Q relativo alla PCA con 3 componenti si osserva che la maggior parte dei campioni è concentrata nella regione in basso a sinistra: questi vini sono ben descritti dal modello e non presentano varianze residue significative. Sono invece presenti alcuni campioni, in particolare appartenenti alla classe GR, quindi dei vini Grignolino, che presentano valori elevati di T^2 e/o di Q . Questi possono essere considerati campioni estremi (e potenzialmente outliers) rispetto alla struttura media dei dati. Ciò indica che, pur avendo catturato la maggior parte dell'informazione con 3 PC, esistono alcune bottiglie di Grignolino con un profilo chimico particolare e distante dalla popolazione globale. Possiamo notare come per i vini di classe OLO, quindi "Barolo" non vi siano quasi campioni discostanti mentre per la categoria dei ERA, ovvero "Barbera" vi sia qualche campione ma molto meno significativi rispetto al "Grignolino".

3

Conclusioni

All'interno di questa analisi si è visto come utilizzando 3 componenti PCA per la creazione di un modello rappresentativo del nostro dataset siamo riusciti a catturare un livello significativo di varianza che ci ha permesso di effettuare le seguenti osservazioni:

- Il pre processing corretto da scegliere in casi di gradienze fisiche diverse è l'autoscaling
- Le variabili che maggiormente influenzano e contribuiscono alle componenti principali utilizzate sono: Nonflav Phen e Hue.
- Tramite grafico degli score è possibile denotare 3 gruppi di campioni e vedere come si posizionano rispetto ai loadings.
- Si è notato inoltre come questi gruppi fossero dei veri e propri cluster appartenendo alle stesse famiglie di vini.
- Infine tramite l'analisi dei residui si è notato come il modello PCA proposto è in grado di catturare sufficientemente l'informazione del dataset mostrando un basso numero di campioni considerabili "estremi" o "outliers".